Sample Size Considerations in Multivariate Normal Classification

by

Seymour Geisser[1] and Wesley Johnson
University of Minnesota and University of California at Davis

Technical Report 580
July 1992

[1]Research supported in part by NIGMS Grant 25271

Sample Size Considerations in Multivariate Normal Classification

by

Seymour Geisser, University of Minnesota
Wesley Johnson, University of California at Davis

1. Introduction

In classification problems involving two multivariate normal training samples of size N_1 and N_2 already in hand we address the question of whether it would be worthwhile to increase the training sample by given amounts. Consider a situation where it is possible that a total of, say n_1 and n_2 observations would be taken from the two populations respectively. Suppose a decision is made to observe $N_1 < n_1$ and $N_2 < n_2$ observations now with the possibility of observing more observations in the future. If it is expected that the full n_1 and n_2 observations will be taken, analysis of the data at this point is termed interim analysis.

Our main focus is on the performance of linear allocation rules; performance is measured by the magnitude of mis-allocation probabilities. At the "interim" stage, we can assess predictive probabilities about various probabilities of error at the "end" of the experiment. For example, the "actual" error rate for Fisher's sample linear discriminant can be estimated at the interim stage and at the end of the experiment. At the interim stage, it may be of interest to assess the chances that this error rate will be less than or greater than .01, .05, .2 etc. after more observations are taken. If it is assessed that the estimated "actual" error rate will be greater than .2 at the "end" of the experiment with predictive probability .99, these may be grounds to terminate the experiment at the interim stage or perhaps to consider additional variables that might aid in lowering the error rate. On the other extreme, if the "actual" error rate is estimated to be less than .01 at the interim stage, and if the predictive probability that it will remain that low is high, it may be deemed unnecessary to observe more data or perhaps continue the experiment

with enthusiasm. Similar considerations will be made with respect to the "true" error rate defined for the population linear discriminant.

The approach taken is Bayesian. Since the problem involves the Mahalanobis measure of divergence D^2, which crops up in testing the similarity of two multivariate normal populations, we initially discuss this problem in sections 2 and 3. Section 4 considers the effect of a potential training sample increase on the "true" errors of classification. The effect of the training sample increase on the "actual" errors of classification is addressed in section 5. The results are exhibited by an example presented in section 6.

2. The Two-Sample Case

For two d-variate normal populations where Π_1 is $N(\mu_1,\Sigma)$ and Π_2, is $N(\mu_2,\Sigma)$, Mahalanobis (1936) introduced a measure of divergence, D^2, between the two populations. However, in keeping with the tradition of using Greek letters for parametric functions we shall denote this measure as

$$\gamma = (\mu_1-\mu_2)' \Sigma^{-1}(\mu_1-\mu_2). \qquad (2.1)$$

In most instances the parameters that constitute γ are unknown and samples of size N_i from Π_i, i=1,2 are used to estimate them or test hypotheses about them. For example, when the random samples yield sample means $\bar{x}_1$, $\bar{x}_2$ and pooled sample covariance matrix S where νS is Wishart, $W(\Sigma,\nu)$, $\nu = N_1+N_2-2$, and $c = N_1N_2/(N_1+N_2)$, then Hotelling's (1931)

$$T^2 = c(\bar{x}_1-\bar{x}_2)' S^{-1}(\bar{x}_1-\bar{x}_2) \qquad (2.2)$$

is used for testing the null hypothesis $\mu_1=\mu_2$. When $\mu_1=\mu_2$, the sampling distribution of T^2 is $\nu d(\nu-d-1)^{-1}$ F(d,ν–d–1) where F(a,b) is the F distribution with a and b degrees of freedom. Under the alternative $\mu_1\neq\mu_2$ the distribution of T^2 is $\nu d(\nu-d-1)^{-1}$F(d,ν–d–1, $c\gamma$) where F(a,b,λ) is a non-central F with non-centrality parameter λ.

Another situation where γ is of interest is when classification of a new observation is at issue. This observation is assumed to have originated from Π_i with probability q_i. When the parameters of Π_i are known the usual classification scheme is to use the population linear discriminant

$$U = \left[z - \frac{1}{2}(\mu_1+\mu_2)'\right]\Sigma^{-1}(\mu_1-\mu_2) \qquad (2.3)$$

for assigning z. The procedure is to assign z to Π_1 if

$$U > \log\frac{q_2}{q_1} \qquad (2.4)$$

and to Π_2 otherwise. Of course U can only be used directly if all the parameters are known. When the population parameters are to be estimated from training samples of size N_1 and N_2, z is assigned to Π_1 if

$$V > \log\frac{q_2}{q_1} \qquad (2.5)$$

and to Π_2 otherwise where

$$V = \left[z - \frac{1}{2}(\bar{x}_1+\bar{x}_2)'\right] S^{-1}(\bar{x}_1-\bar{x}_2) \qquad (2.6)$$

is the sample linear discriminant that is used as an estimate of U. This of course is the one actually used. Now the distribution of U conditional on μ_1, μ_2 and Σ is $N(\frac{1}{2}\gamma,\gamma)$ under Π_1 and $N(-\frac{1}{2}\gamma,\gamma)$ under Π_2.

Other applications where γ is of interest are in profile analysis and cluster analysis, Pillai (1985). Hence the estimation of γ is of interest in a variety of problems involving samples from multivariate populations. As in Geisser (1967) we shall consider the problem from a Bayesian viewpoint. There it was assumed that in the absence of prior knowledge of μ_1, μ_2 and Σ that it might be reasonable to assign the prior density

$$p(\mu_1,\mu_2,\Sigma^{-1}) \propto |\Sigma|^{(d+1)/2}. \tag{2.7}$$

It was also shown there that the posterior density of γ is

$$p_{T^2}(\gamma) = \sum_{j=0}^{\infty} \frac{c(c\gamma)^{\frac{1}{2}(d+2j)-1} e^{-c\gamma/2} \Gamma\left(\frac{\nu}{2}+j\right)\left(\frac{\nu}{T^2}\right)^{\nu/2}}{2^{d/2+j}\,\Gamma[(d+2j)/2]\,\Gamma(\nu/2)\left(1+\frac{\nu}{T^2}\right)^{\frac{\nu}{2}+j} j!} \tag{2.8}$$

or

$$p_{T^2}(c\gamma) = \sum_{j=0}^{\infty} w_j\, f(c\gamma \mid d+2j) \tag{2.9}$$

where the coefficients

$$w_j = \binom{\nu/2+j-1}{j}\left(\frac{\nu}{\nu+T^2}\right)^{\nu/2}\left(\frac{T^2}{\nu+T^2}\right)^{j} \tag{2.10}$$

are negative binomial weights and $f(\cdot|d+2j)$ is the density of a χ^2 with $d+2j$ degrees of freedom. Hence for cases involving tests

$$H_o: \gamma \leq \gamma_o \quad \text{vs.} \quad H_1: \gamma > \gamma_o$$

one can base a test of H_o on whether the posterior probability

$$\Pr[\gamma > \gamma_o \mid x^{(N_1+N_2)}] \geq p \tag{2.11}$$

for some specified p, $0 < p < 1$, where $x^{(N_1+N_2)}$ represents all of the data in the samples of size N_1 and N_2 respectively.

Now an interesting problem that often occurs is when considerable importance is attached to H_1 and the samples in hand have not attained the value p considered necessary to conclude that H_1 is appropriate. At this point one can construct additional "thought samples" of size M_1 and M_2 and calculate whether the chance of achieving the desired result is sufficiently large to take the additional samples. Such a procedure was already devised for a single sample and a given distance, Geisser and Johnson (1992). Here we will first complete the analysis for the two sample case. Secondly we use these results to make the appropriate calculations for the classification problem in order to ascertain what additional sample sizes might be necessary to achieve particular goals with regard to classification errors.

3. Predicting the Rejection of H_o

Now we calculate the predictive probability P of rejecting H_o if M_1 and M_2 additional observations were sampled from Π_1 and Π_2 respectively. Let

$$P = Pr\left[Pr\left[\gamma > \gamma_0 \mid x_1^{(N_1)}, x_2^{(N_2)}, X_{(M_1)}, X_{(M_2)} \right] \geq p \right] \tag{3.1}$$

for the observed training samples $x_i^{(N_i)} = (x_{i1},...,x_{iN_i})$ and the potential ones,

$X_{(M_i)} = (X_{i1},...X_{iM_i})$ from Π_i, i=1,2.

Setting $N_1 + M_1 = n_1$, $N_2 + M_2 = n_2$, to reject H_o we require

$$Pr\left[c_1\gamma > c_1\gamma_0 \mid x_1^{(n_1)}, x_2^{(n_2)} \right] \geq p \tag{3.2}$$

for $c_1 = n_1n_2/(n_1+n_2)$, $\nu_1 = n_1+n_2-2$. We make use of the distribution of γ with appropriate substitutions. Now (3.2) is equivalent to

$$1 - F_{T2}(c_1\gamma_0) \geq p \tag{3.3}$$

where $F_{T2}(c_1\gamma)$ is the distribution function of $c_1\gamma$ whose density is given in (2.9) with the appropriate substitutions n_i for N_i, c_1 for c and ν_1 for ν. Therefore if we have only observed samples of size N_1 and N_2 and wish to predict whether additional samples of size M_1 and M_2 would reject H_o we need to calculate

$$P = Pr[1 - F_{T2}(c_1\gamma_0) \geq p]. \tag{3.4}$$

It can be shown, similarly as in Geisser and Johnson (1992), that $1 - F_{T2}(c_1\gamma_0)$ is increasing in T^2. Hence we need only find the minimum T^2 say t_0^2 such that

$$1 - F_{t_o^2}(c_1\gamma_0) \geq p \tag{3.5}$$

and then

$$P = Pr(T^2 \geq t_o^2),$$

for

$$T^2 = c_1(\bar{y}_1 - \bar{y}_2)' S^{-1}_{n_1+n_2} (\bar{y}_1 - \bar{y}_2),\ c_1 = \frac{(N_1+M_1)(N_2+M_2)}{N_1+N_2+M_1+M_2} \tag{3.6}$$

where $S_{n_1+n_2}$ is the pooled sample covariance matrix based on n_1+n_2 observations and

$$\bar{y}_i = \frac{N_i\bar{x}_{N_i} + M_i\bar{X}_{M_i}}{N_i + M_i},\ i=1,2.$$

As in the one sample case treated in Geisser and Johnson (1992) we shall approximate $S^{-1}_{n_1+n_2}$ by $S^{-1}_{N_1+N_2}$ to obtain a more tractable result. Define

$$\hat{T}^2 = c_1(\bar{y}_1 - \bar{y}_2)' S^{-1}_{N_1+N_2} (\bar{y}_1 - \bar{y}_2) \tag{3.7}$$

and

$$D = (\bar{x}_{N_1} - \bar{x}_{N_2})' Q^{-1}(\bar{x}_{N_1} - \bar{x}_{N_2})/q \tag{3.8}$$

where

$$Q = \left(\frac{M_1}{N_1(N_1+M_1)} + \frac{M_2}{N_2(N_2+M_2)} \right) S_{N_1+N_2}.$$

Then in a manner similar to the previously mentioned one sample case we find that for $q = N_1+N_2-d-1$

$$B = \frac{\hat{T}^2 N_1 N_2 (N_1+M_1+N_2+M_2)}{q(1+D)[M_1 N_2 (N_2+M_2) + M_2 N_1 (N_1+M_1)]} \tag{3.9}$$

has predictive density

$$f(b) = \sum w_k \, f(b \mid k+\frac{d}{2}, k+\frac{q}{2})$$

where

$$f\left(b \mid k+\frac{d}{2}, k+\frac{q}{2}\right) \propto b^{k+\frac{d}{2}-1} (1+b)^{-\left(2k+\frac{d+q}{2}\right)} \tag{3.10}$$

with negative binomial coefficients,

$$w_k = \binom{k+\frac{q}{2}-1}{k} \left(\frac{D}{1+D}\right)^k \left(\frac{1}{1+D}\right)^{q/2},$$

where $\bar{X}_{M_i}$, i=1,2 are treated as yet unobserved random variables. Hence

$$P = Pr(T^2 \geq t_o^2) \doteq P(\hat{T}^2 \geq t_o^2) \tag{3.11}$$

or

$$P \doteq \Pr\left(B > \frac{t_o^2 N_1 N_2 (N_1+M_1+N_2+M_2)}{q(1+D)[M_1 N_2 (N_2+M_2)+M_2 N_1 (N_1+M_1)]} \right)$$

which can be calculated to reasonable accuracy.

4. The Expected Effect of Increasing the Training Samples on the "True" Errors of Classification

Now the "true" errors of classification inherent in the set of variables used for classifying a future observable Z are

$$\varepsilon_1 = \Pr\left[U < \log \frac{q_2}{q_1} \,\middle|\, \mu_1, \mu_2, \Sigma, Z \in \Pi_1 \right] = \Phi(\tau_1) \tag{4.1}$$

$$\varepsilon_2 = \Pr\left[U > \log \frac{q_2}{q_1} \,\middle|\, \mu_1, \mu_2, \Sigma, Z \in \Pi_2 \right] = 1 - \Phi(\tau_2) \tag{4.2}$$

where $\Phi(\cdot)$ is the standard normal distribution function,

$$\tau_1 = \frac{\left(\log \frac{q_2}{q_1} - \frac{1}{2}\gamma \right)}{\gamma^{\frac{1}{2}}} \tag{4.3}$$

$$\tau_2 = \frac{\left(\log \frac{q_2}{q_1} + \frac{1}{2}\gamma \right)}{\gamma^{\frac{1}{2}}}, \tag{4.4}$$

and U is defined in (2.3). Suppose $q_2 \geq q_1$. (This is always possible by relabelling the populations.) Then ε_1 is a monotone decreasing function of γ. Now for a given $r < \frac{1}{2}$

$$\Pr[\varepsilon_1 \le r] = \Pr[\varepsilon_1 \le \Phi(\tau_{1r})] \tag{4.5}$$

where

$$r = \Phi(\tau_{1r}) = \Pr[\tau_1 \le \tau_{1r}] \quad \text{and} \quad \tau_{1r} < 0.$$

Then

$$\Pr[\varepsilon_1 \le r] = \Pr\left[\frac{1}{2}\gamma + \tau_{1r}\gamma^{\frac{1}{2}} > \log\frac{q_2}{q_1}\right] \tag{4.6}$$

which can be numerically calculated using the density $p_{T2}(\gamma)$ given in (2.8). For $q_2 = q_1$ (4.6) simply reduces to

$$\Pr[\varepsilon_1 \le r] = \Pr[\gamma > 4\tau_{1r}^2] \tag{4.7}$$

recalling that $-2\tau_{1r} > 0$. In this case a similar result for ε_2 yields

$$\Pr[\varepsilon_2 \le r'] = \Pr[\gamma > 4\tau_{1r'}^2]. \tag{4.8}$$

Suppose we anticipate adding M_1 and M_2 additional observations to the two training samples from Π_1 and Π_2. We then would be interested in calculating the predictive probabilities

$$P = \Pr[\Pr[\varepsilon_1 \le r] \ge p] \tag{4.9}$$

and

$$P' = Pr[Pr[\varepsilon_2 \le r'] \ge p']. \tag{4.10}$$

For the case $q_2 = q_1$ the results simplify to calculating

$$P = Pr\left[Pr[\gamma > 4\tau_{1r}^2] \ge p\right] \tag{4.11}$$

$$P' = Pr\left[Pr[\gamma > 4\tau_{1r'}^2] \ge p'\right]. \tag{4.12}$$

Using the approximate result in the previous section

$$P \doteq Pr\left[B > \frac{t_{op}^2 N_1 N_2 (N_1+M_1+N_2+M_2)}{q(1+D)[M_1 N_2 (N_2+M_2)+M_2 N_1 (N_1+M_1)]}\right] \tag{4.13}$$

where t_{op}^2 is defined as in (3.5) i.e. the minimum $\hat{T}^2$ such that

$$1 - F_{t_{op}^2}(c_1 \gamma_o) \ge p \tag{4.14}$$

and $\gamma_o = 4\tau_{1r}^2$. A similar calculation is made for P' with appropriate substitutions of p', $t_{op'}^2$ and $\tau_{1r'}^2$.

These calculations would provide guidance on whether it would be worthwhile to increase the training samples by the proposed amounts given the stated probabilistic criteria for the estimation of the "true" errors of classification.

5. The Effect on "Actual" Errors of Classification

The sample linear discriminant V of (2.6) will now be investigated with regard to the potential increments in the training sample previously discussed.

Attention is directed to the "actual" errors of classification, namely the predictive probabilities of misclassification when the sample discriminant V is used. They are

$$e_1 = E\left\{ \Pr\left[V < \log \frac{q_2}{q_1} \,\middle|\, z \in \Pi_1, \mu_1, \mu_2, \Sigma^{-1} \right] \right\} \tag{5.1}$$

$$e_2 = E\left\{ \Pr\left[V > \log \frac{q_2}{q_1} \,\middle|\, z \in \Pi_2, \mu_1, \mu_2, \Sigma^{-1} \right] \right\} \tag{5.2}$$

where the expectation is over the posterior distribution of $\mu_1, \mu_2, \Sigma^{-1}$ given $x_1^{(n_1)}, x_2^{(n_2)}$.

Now with $s < \frac{1}{2}$ and $s' < \frac{1}{2}$, we need to calculate the predictive probabilities

$$\Pr[e_1 \le s] \text{ and } \Pr[e_2 \le s'] \tag{5.3}$$

where e_1 and e_2 are random since V is a function of $\bar{y}_i = \dfrac{N_i \bar{x}_{iN_i} + M_i \bar{X}_{iM_i}}{N_i + M_i}$, i=1,2.

Now as in Geisser (1967)

$$e_1 = \Pr\left[t_{\nu_1+1-d} < \left(\log \frac{q_2}{q_1} - \frac{T^2}{2c_1} \right) \left[\frac{c_1 n_1 (\nu_1+1-d)}{\nu_1 (n_1+1) T^2} \right]^{\frac{1}{2}} \right] \tag{5.4}$$

$$e_2 = \Pr\left[t_{\nu_1+1-d} > \left(\log \frac{q_2}{q_1} + \frac{T^2}{2c_1} \right) \left[\frac{c_1 n_2 (\nu_1+1-d)}{\nu_1 (n_2+1) T^2} \right]^{\frac{1}{2}} \right] \tag{5.5}$$

where $\nu_1 = n_1+n_2-2$ and t_{ν_1+1-d} is a student t variate with ν_1+1-d degrees of freedom. Now as T^2 increases, both e_1 and e_2 decrease. Hence we need to find the minimum T^2 such that

$$Pr[e_1 \leq s] = Pr[T^2 \geq t^2_{os}] \tag{5.6}$$

where t^2_{os} is the minimum value of T^2 such that $e_1 \leq s$ and

$$Pr[e_2 \leq s'] = Pr[T^2 \geq t^2_{os'}] \tag{5.7}$$

where $t^2_{os'}$ is the minimum value of t^2 such that $e_2 \leq s'$. Now as before because of the complexity of the distribution of T^2 we use $\hat{T}^2$ as its approximation and

$$P = Pr[e_1 \leq s] \doteq Pr[\hat{T}^2 \geq t^2_{os}] \tag{5.8}$$

$$P' = Pr[e_2 \leq s'] \doteq Pr[\hat{T}^2 \geq t^2_{os'}] \tag{5.9}$$

where the distribution of $\hat{T}^2$ is given as in (3.11).

Since q_i was assumed known one can maximize a function of P and P' for a fixed total of the prospective training samples M_1+M_2 by examining all variations in potential allotments of M_1. When q_i is estimated from the training samples i.e. a random sample of $N = N_1+N_2$ is taken from $\Pi = (\Pi_1,\Pi_2)$ and N_1 and N_2 then determined, the q_i's can then be estimated in a Bayesian fashion Geisser (1964). Hence there are two alternatives that can be envisaged. First fixed samples of size M_1 and M_2. This case is easily managed much as before. A second alternative is when the future training sample of size M is also presumed to be drawn from Π, then one can then modify the calculations based on the future expectations of M_1 and M_2 for the fixed $M = M_1+M_2$.

6. Illustrative Example

As an illustration we consider a subset of the Iris data of Fisher (1936). The full set consists of 4 variables yielding measurements on sepal and petal widths and lengths in centimeters on 150 plants, 50 each from 3 different species of irises, setosa, versicolor and virginica. For the sake of illustration, we only consider two of the variables, sepal and petal width on each plant, and 2 species - versicolor and virginica. Further we shall conduct an interim analysis using a sample of 25 sets of observations in Fisher's table and assume $q_1 = q_2$. Hence $N_1 = N_2 = 25$ and assume $M_1 = M_2 = 25$. Table 1 lists the posterior probabilities at the interim stage that a future versicolor observation will be incorrectly allotted to virginica and vice versa because the errors will be the same for this situation.

(Table 1 about here)

If our goal were to see if 25 additional observations of both species could drive the true error below .05 then from the table it is clear that there is only a slight chance for $p = .8$ and virtually no chance for $p \geq .9$. On the other hand, the chance of driving ε_i below .1 with the additional 50 observations is rather high for $p \leq .95$. We also see that with virtual predictive certainty that $\varepsilon_i \leq .2$ if the anticipated sample were taken, even with $p = .995$. In this case the extra 50 observations are actually there to use and the calculation appears in the table for $N_i = 50$ indicating that our predictions about them were supported.

We now consider the "actual" errors of classification for $N_i = M_i = 25$, $q_i = \frac{1}{2}$, i=1,2. The results are displayed in Table 2.

(Table 2 about here)

We note that the interim actual error at $N_i = 25$ was .052 and that the anticipated next 25 observations would have made it highly likely, $P = .89$, that e_i at $M_i = 25$ would remain below .06. It in fact did so but barely as e_i at the actual $N_i = 50$ was .059. On the other hand at $N_i = 25$ the chance was .18 that at $M_i = 25$ the error rate would not exceed

.04. It would appear then if an actual error rate of about .05 were acceptable, the additional training samples would not be necessary to obtain unless there was no cost involved. Further the actual error at $N_i = 25$ is obviously rather close to the expected true error whose lower bound is $\Phi[-\frac{1}{2}(c^{-1}d+c^{-1}T^2)^{1/2}] = .042$ which is negligibly smaller than the actual error of .052. For $N_i = 50$, the lower bound on the expected true error is .054 and the actual error is .059. For classification purposes it would appear that the training sample of $N_i = N_2 = 25$ would have sufficed from several points of view for purposes of classification.

The sample means and pooled covariance matrices based on $N_1 = N_2 = 25$ and $N_1 = N_2 = 50$ appear in Table 3.

(Table 3 about here)

References

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, Vol. VII, Pt. II, 179-188.

Geisser, S. (1964). Posterior odds for multivariate normal classification. Journal of the Royal Statistical Society B, 1, 69-76.

Geisser, S. (1967). Estimation associated with linear discriminants. Annals of Mathematical Statistics, 38, 3, 807-817.

Geisser, S. and Johnson, W.O. (1992). Interim analysis for normally distributed observables. Proceedings of the International Symposium on Multivariate Analysis and Its Application. (To appear.)

Hotelling, H. (1931). The generalization of student's ratio. Annals of Mathematical Statistics, 2, 360-378.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. Proceedings National Institute of Science, India 12, 49-55.

Pillai, K.C. (1985). Mahalanobis D^2, Encyclopedia of Statistical Sciences, 5, 176-181.

Table 1. Interim probabilities at $N_i = 25$ and $N_i = 50$, and the approximate predictive probabilities P of the true error for $N_i = 25$, $M_i = 25$, $q_i = \frac{1}{2}$, i=1,2 for several values of r and p.

r	$Pr[\varepsilon_i \leq r]$		Approximate P				
	$N_i = 25$,	$N_i = 50$	p = .8	.9	.95	.99	.995
.01	.003	*	*	*	*	*	*
.02	.06	.002	*	*	*	*	*
.03	.22	.03	.0003	*	*	*	*
.04	.43	.16	.01	.0007	4×10^{-5}	*	*
.05	.61	.39	.13	.02	.002	*	*
.06	.76	.63	.46	.14	.03	.0002	3×10^{-5}
.07	.85	.79	.76	.40	.15	.003	.0005
.08	.92	.90	.93	.71	.40	.03	.006
.09	.95	.96	.98	.90	.69	.12	.04
.10	.97	.98	.997	.97	.88	.34	.17
.20	**	**	**	**	**	**	**

* indicates entry $< 10^{-5}$

** indicates entry > .9999

Table 2. Interim probability e_i for $N_i = 25$ and the approximate predictive probability P for $M_i = 25$ for "actual" error for several values of s and p with $q_i = \frac{1}{2}$, and e_i for $N_i = 50$.

s	.01	.02	.03	.04	.05	.06	.07	.08	.09	.1	.2
Approx. P	*	7×10^{-5}	.01	.18	.58	.89	.98	.996	.999	**	**

$e_i = .052$ at $N_i = 25$

$e_i = .059$ at $N_i = 50$

* indicates entry $< 10^{-5}$

** indicates entry $> .9999$

Table 3. Sample Means and Covariance Matrices for $N_i = 25, 50$

	Versicolor (Sample Means)		Virginica	
	Petal Width	Sepal Width	Petal Width	Sepal Width
$N_i = 25$	2.776	1.344	2.936	2.076
$N_i = 50$	2.770	1.326	2.974	2.026

$$S_{50} = \begin{pmatrix} .1034 & .0483 \\ .0483 & .0602 \end{pmatrix}, \quad S_{100} = \begin{pmatrix} .1012 & .0444 \\ .0444 & .0573 \end{pmatrix}$$